DOCUMENT RESUME

ED 459 821 IR 058 357

AUTHOR Myers, Brad A.; Casares, Juan P.; Stevens, Scott; Dabbish, Laura; Yocum, Dan; Corbett, Albert

TITLE A Multi-View Intelligent Editor for Digital Video Libraries.

SPONS AGENCY National Science Foundation, Arlington, VA.

PUB DATE 2001-06-00

NOTE 11p.; In: Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (1st, Roanoke, Virginia, June 24-28, 2001). For entire proceedings, see IR 058 348. Figures may not reproduce well.

CONTRACT IIS-9817527

AVAILABLE FROM Association for Computing Machinery, 1515 Broadway, New York NY 10036. Tel: 800-342-6626 (Toll Free); Tel: 212-626-0500; e-mail: acmhelp@acm.org. For full text: http://www1.acm.org/pubs/contents/proceedings/dl/379437/.

PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS Audiovisual Aids; Authoring Aids (Programming); *Editing; Electronic Libraries; Information Technology; *Interactive Video; Multimedia Materials; Video Equipment

IDENTIFIERS *Digital Technology; Metadata

ABSTRACT

Digital video is becoming increasingly ubiquitous. Most camcorders today are digital, and computers are being advertised based on their video editing capabilities. Many exciting research projects are investigating how to search, visualize, and summarize digital video, but there is little work on new ways to support the use of the video beyond just playing it. In fact, editing video is significantly more difficult than editing textual material. Silver (Simplifying Interactive Layout and Video Editing and Reuse) is an authoring tool that aims to allow novice users to edit digital video. The goal is to make editing of digital video as easy as text editing. Silver provides multiple coordinated views, including project, source, outline, subject, storyboard, textual transcript and timeline views. Selections and edits in any view are synchronized with all other views. A variety of recognition algorithms are applied to the video and audio content and then are used to aid in the editing tasks. The Informedia Digital Library supplies the recognition algorithms and metadata used to support intelligent editing, and Informedia also provides search and a repository. The metadata includes shot boundaries and a time-synchronized transcript, which are used to support intelligent selection and intelligent cut/copy/paste. (Contains 35 references.) (Author/AEF)

Reproductions supplied by EDRS are the best that can be made
from the original document.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. Cotton

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

# A Multi-View Intelligent Editor for Digital Video Libraries

Brad A. Myers, Juan P. Casares, Scott Stevens, Laura Dabbish, Dan Yocum, Albert Corbett

Human Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
bam@cs.cmu.edu
http://www.cs.cmu.edu/~silver

ED 459 821

## ABSTRACT

Silver is an authoring tool that aims to allow novice users to edit digital video. The goal is to make editing of digital video as easy as text editing. Silver provides multiple coordinated views, including project, source, outline, subject, storyboard, textual transcript and timeline views. Selections and edits in any view are synchronized with all other views. A variety of recognition algorithms are applied to the video and audio content and then are used to aid in the editing tasks. The Informedia Digital Library supplies the recognition algorithms and metadata used to support intelligent editing, and Informedia also provides search and a repository. The metadata includes shot boundaries and a time-synchronized transcript, which are used to support intelligent selection and intelligent cut/copy/paste.

### Keywords

Digital video editing, multimedia authoring, video library, Silver, Informedia.

## 1. INTRODUCTION

Digital video is becoming increasingly ubiquitous. Most camcorders today are digital, and computers are being advertised based on their video editing capabilities. For example, Apple claims that you can "turn your DV iMac into a personal movie studio" [1]. There is an increasing amount of video material available on the World-Wide-Web and in digital libraries. Many exciting research projects are investigating how to search, visualize, and summarize digital video, but there is little work on new ways to support the use of the video beyond just playing it. In fact, editing video is significantly harder than editing textual material. To construct a report or a new composition using video found in a digital library or using newly shot video requires considerably more effort and time than creating a similar report or composition using quoted or newly authored text.

In the Silver project, we are working to address this problem by bringing to video editing many of the capabilities long available in textual editors such as Microsoft Word. We are also trying to alleviate some of the special problems of video editing. In particular, the Silver video editor provides multiple views, it uses the familiar interaction techniques from text editing, and it provides intelligent techniques to make selection and editing easier.

The Silver editor is designed to support all phases of the video post-production process. The storyboard and script views support brainstorming and planning for the video. The project, source, subject and outline views support the collection and organization of the source material. The timeline and script views support the detailed editing, and the preview view can be used to show the result.

Silver is an acronym and stands for Simplifying Interactive Layout and Video Editing and Reuse. The key innovations in the Silver editor include: providing a transcript view for the actual audio; multiple views with coordinated selections, including the ability to show when one view only contains part of the selection; intelligent context-dependent expansion of the selection for double-clicking; and intelligent cut/copy/paste across the video and audio. These are discussed in this article.

## 2. STATE OF THE ART

Most tools for editing video still resemble analog professional video editing consoles. Although they support the creation of high quality material, they are not easy for the casual user, especially when compared with applications such as text editors. Current video editing software only operates at a low syntactic level, manipulating video as a sequence of frames and streams of uninterpreted audio. It does not take advantage of the content or structure of the video or audio to assist in the editing. Instead, users are required to pinpoint specific frames, which may involve zooming and numerous repetitions of fast-forward and rewind operations. In text editing, the user can search and select the content by letters, words, lines, sentences, paragraphs, sections, etc. In today's video and audio editing, the only units are low-level frames or fractions of seconds.

As another example, three-point editing is one option in professional editors such as Adobe Premiere. In this kind of editing, the user can locate an "in point" and an "out point"

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
*JCDL'01*, June 24-28, 2001, Roanoke, Virginia, USA.
Copyright 2001 ACM 1-58113-345-6/01/0006...$5.00.

IR058357



BEST COPY AVAILABLE

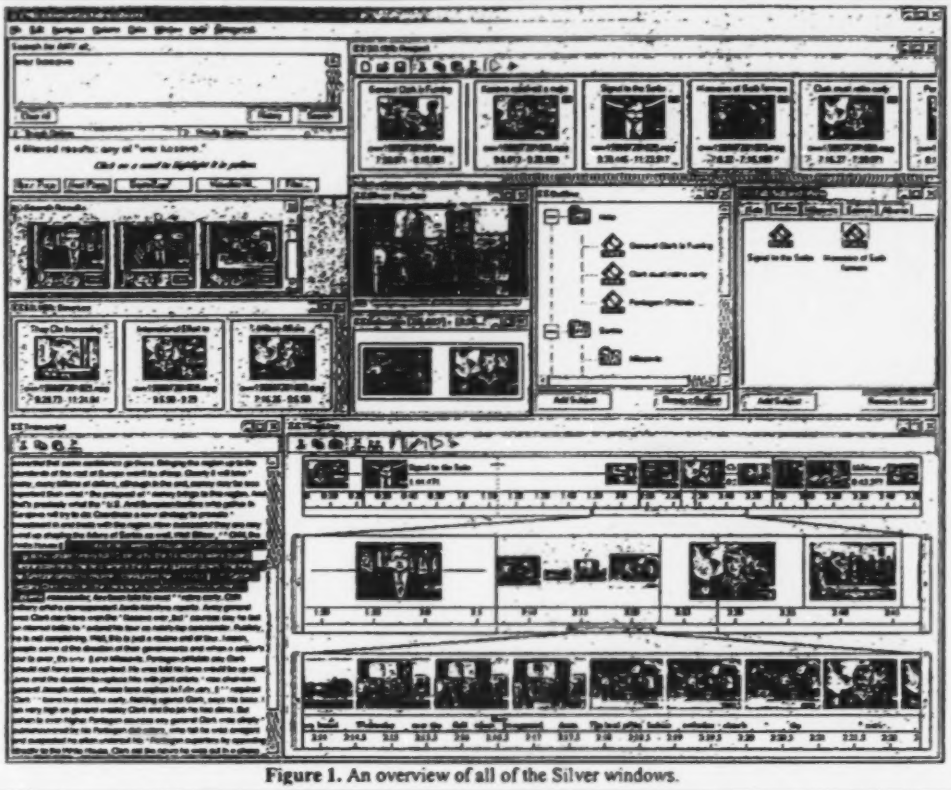

**Figure 1.** An overview of all of the Silver windows.

in the video source and a third point in the target to perform a copy operation. The fourth point for the edit is computed based on the length of the in-out pair. This method can be traced to the use of physical videotape where the length of the output and input segments must be the same. However, three-point editing is not required for digital video and is very different from the conventional cut/copy/paste/delete technique that is used in other editing tools on computers.

## 3. RELATED WORK

There is a large body of work on the extraction and visualization of information from digital video (e.g., [29] [12]). However, most of this work has focused on automatic content extraction and summarization during library creation and searching, and on information presentation during library exploration. In the Silver project, we focus on authoring with the content once it is found.

After examining the role of digital video in interactive multimedia applications, Mackay and Davenport realized that video could be an information stream that can be tagged, edited, analyzed and annotated [19]. Davenport et. al. proposed using metadata for home-movie editing assistance [10]. However, they assumed this data would be obtained through manual logging or with a "data camera" during filming, unlike the automatic techniques used in Silver.

The Zodiac system [5] employs a branching edit history to organize and navigate design alternatives. It also uses this abstraction to automatically detect shot and scene boundaries and to support the annotation of moving objects in the video. IMPACT [33] uses automatic cut detection and camera motion classification to create a high level description of the structure of the video, and then visualizes and edits the structure using timeline and tree structure views [32]. IMPACT also detects object boundaries and can recognize identical objects in different shots.

The Hierarchical Video Magnifier [24] allows users to work with a video source at fine levels of detail while maintaining an awareness of the context. It provides a timeline to represent the total duration of the video source, and supplies the user with a series of low-resolution frame samples. There is also a tool that can be used to expand or reduce the effective temporal resolution of any portion of the timelines. Successive applications of the temporal magnifier create an explicit spatial hierarchical structure of the video source. The Swim Hierarchical Browser [35] improves on this idea by using automatically detected shots in the higher level layers. These tools were only used for top-down navigation, and not for editing. We use a similar approach in our Timeline view.



BEST COPY AVAILABLE

The Video Retrieval and Sequencing System [8] semiautomatically detects and annotates shots for later retrieval. Then, a cinematic rule-based editing tool sequences the retrieved shots for presentation within a specified time constraint. For example, the parallel rule alternates two different sets of shots and the rhythm rule selects longer shots for a slow rhythm and shorter shots for a fast one.

Most video segmentation algorithms work bottom-up, from the pixels in individual frames. Hampapur [15] proposes a top-down approach, modeling video-editing techniques mathematically to detect cuts, fades and translation effects between shots.

Video Mosaic [20] is an augmented reality system that allows video producers to use paper storyboards as a means of controlling and editing digital video. Silver's storyboards are more powerful since they can also be used for interactive videos. CVEPS [23] automatically extracts key visual features and uses them for browsing, searching and editing. An important contribution of this system is that it works in the compressed domain (MPEG), which has advantages in terms of storage, speed and noiseless editing.

VideoScheme [22] is a direct manipulation video editing system that provides the user with programming capabilities. This increases the flexibility and expressiveness of the system, for example supporting repetitive or conditional operations. VideoMAP [31] indexes video through a variety of image processing techniques, including histograms and "x-rays" (edge pixel counts). The resulting indices can be used to detect cuts and camera operations and to create visualizations of the video. For example, VideoMAP renders the indices over time, and VideoSpaceIcon represents the temporal and spatial characteristics of a shot as an icon.

The Hitchcock system [14] automatically determines the "suitability" of the different segments in raw video, based on camera motion, brightness and duration. Similar clips are grouped into "piles." To create a custom video, the user drags segments into a storyboard, specifies a total desired duration and Hitchcock automatically selects the start and end points of each clip based on shot quality and total duration. Clips in the storyboard are represented with frames that can be arranged in different layouts, such as a "comic book" style layout [2]. We plan to incorporate similar techniques into Silver.

## 4. INFORMEDIA

We obtain our source video and metadata through CMU's Informedia Digital Video Library [34]. The Informedia project is building a searchable multimedia library that currently has over 2,000 hours of material, including documentaries and news broadcasts. Informedia adds about two hours of additional news material every day. For all of its video content, Informedia creates a textual transcript of the audio track using closed-captioning information and speech recognition [7]. The transcript is time-aligned with the video using CMU's Sphinx speech recognition system [27]. Informedia also performs image analysis to detect shot boundaries, extracting representative thumbnail images from each shot [6] and detects and identifies faces in the video frame. A video OCR system identifies and recognizes captions in the image [28]. Certain kinds of camera movements such as pans and fades can also be identified. All of this metadata about the video is stored in a database. This metadata is used by Informedia to automatically create titles, representative frames and summaries for video clips, and to provide searching for query terms and visualization of the results. Silver takes advantage of the metadata in the database to enhance its editing capabilities.

## 5. TYPES OF PRODUCTIONS

The Silver video editor is designed to support different kinds of productions. Our primary goal is to make it easier for middle and high school children (ages 10-18) to create multimedia reports on a particular topic. For example, a social-studies teacher might have students create a report on Kosovo. We want to make it as easy for students to create such a video report using material in the Informedia library, as it would be to use textual and static graphical material from newspaper and magazine articles.

We also want Silver to support original compositions of two types. First, people might just shoot some video with a camcorder, and then later want to edit it into a production. In the second type, there might first be a script and even a set of storyboards, and then video is shot to match the script. In these two cases where new material is shot, we anticipate that the material will be processed by Informedia to supply the metadata that Silver needs. (Unfortunately, Informedia does not yet give users the capability to process their own video, but Silver is being designed so that we are ready when it does.)

Finally, in the future, we plan for Silver to support *interactive* video and other multi-media productions. With an interactive video, the user can determine what happens next, rather than just playing it from beginning to end. For example, clicking on hot spots in "living books" type stories may choose which video is played next. Another type of production is exemplified by the DVD video "What could you do?" [11]. Here, a video segment is played to set up a situation, then the user is asked a question and depending on the user's answer, a different video piece is selected to be played next.

## 6. MULTIPLE VIEWS

In order to support many different kinds and styles of editing, it is useful to have different views of the material. For example, the Microsoft Word text editor supplies an outline view, a "normal" view that is good for editing, a "print layout" view that more closely shows what the composition will look like, and a "print preview" view that

tries to be exactly like the printed output. Similarly, PowerPoint provides outline, normal, notes, slide sorter, and slide show views. However, video editors have many fewer options. Adobe Premiere's principal view is a frame representation, with a thumbnail image representing one or more video frames. Premiere also provides a Project window, a Timeline window, and a Monitor window (to playback video). MGI's VideoWave III editor has a Library window (similar to the project view), a StoryLine window (a simplified timeline), and a View Screen window for playback. Apple's Final Cut Pro provides a Browser window (like a project view that allows clip organization), Timeline view, Canvas (editing palette), and Viewer window for playback.

Silver currently provides nine different views: the Informedia search and search results, project, source, subject, outline, storyboard, transcript, timeline and preview views. These are described below. Each appears in its own window, which the user or system can arrange or hide if not needed (Figure 1 shows an overview of the full screen).

Figure 2. Search and search results views from Informedia.

### 6.1 Search Results View

When starting from a search using Informedia, the search results will appear in an Informedia search results window (see Figure 2). Informedia identifies *clips* of video that are relevant to the search term, and shows each clip with a representative frame. Each clip will typically contain many different scenes, and the length of a clip varies from around 40 seconds to four minutes. The user can get longer segments of video around a clip by moving up a level to the full video. If the user makes a new query, then the search results window will be erased and the new results will appear instead. Clicking on the thumbnail image will show the clip using the Windows Media Player window. Clicking on the filmstrip icon displays thumbnail images from each shot in the scene, giving a static, but holistic view of the entire clip.

### 6.2 Source and Project Views

When the user finds appropriate video by searching in Informedia, the clips can be dragged and dropped into Silver's Project View (actually, a clip can be dragged directly to any other view, and it will be automatically added to the project view). The Project View (see Figure 3) will also allow other video, audio and still pictures from the disk or the World-Wide Web to be loaded and made easily available for use in the production. As in other video editors such as Premiere, the project view is a "staging area" where the video to be used in a production can be kept. However, in the Silver project view, video clips that are in the composition are separated by a line from clips that are not currently in use. Dragging a clip across this line adds or removes it from the composition.

The source view is like a simplified project view, and allows easy access to the original sources of video. Clips in this view represent the video as brought in from Informedia, from a file, or from the web. They cannot be edited, but they may be dragged to other windows to add a copy of the source to the project.

Figure 3. Silver's project view collects the source material to be used in the production. The first clip is shown on the screen outlined in light yellow to indicate it is partially selected.

### 6.3 Transcript View

An important innovation in the Silver video editor, enabled by Informedia, is the provision of a textual transcript of the video. This is displayed in a conventional text-editor-like window (see Figure 4). Informedia generates the transcript from various sources [16]. If the video provides closed captioning, then this is used. Speech recognition is used to recognize other speech, and also to align the transcript with the place in the audio track where each word appears [7]. Because the recognition can contain mistakes, Silver inserts green "•"s where there appears to be gaps, misalignments, or silence. In the future, we plan to allow the user to correct the transcript by typing the correct words, and then use the speech recognizer to match the timing of the words to the audio track.

The transcript view and the timeline view (section 6.4) are the main ways to specify the actual video segments that go into the composition. In the transcript view, the boundary between segments is shown as a blue double bar ("‖"). The transcript and timeline views are used to find the desired portion of each clip. Transcripts will also be useful in supporting an easy way to search the video for specific content words.



BEST COPY AVAILABLE

We also plan to use the transcript view to support the authoring of new productions from scripts. The user could type or import a new script, and then later the system would automatically match the script to the audio as it is shot.

### 6.4 Timeline View

In Silver, like many other video editors such as Premiere, the Timeline view is the main view used for detailed editing. In order to make the Timeline view more useful and easier to use, we are investigating some novel formats. As shown in Figure 5, we are currently providing a three-level view.

This allows the user to work at a high level of detail without losing the context within the composition. The top level always represents the entire video. The topmost row displays the clips in the composition and their boundaries. For each clip, it shows a representative frame, and if there is enough space, the end frame, title and duration. Below the top row are the time codes. At the bottom of the top level is an indicator showing what portions are being viewed in the bottom two levels. The purple portion is visible in the middle level, and the cyan portion is visible in the bottom level.

The middle level displays the individual shots, as detected by Informedia. Shot boundaries are detected by a change in the video [17]. Each shot is visualized using the representative frame for the shot as chosen by Informedia. The size of the frame is proportional to the duration of the shot.

The bottom level can display the individual frames of the video, so the user can quickly get to particular cut points. The middle row of the bottom level represents the transcript. The bottom level also provides the ability to add annotations or comments to the video (see Figure 6).

A key feature of the Silver timeline is that it allows different representations to be shown together, allowing the user to see the clip boundaries, the time, samples of frames, the cuts, the transcript, annotations, etc. Later, using facilities already provided by Informedia, we can add labels for recognized faces and the waveform for the audio to the timeline. Snapping and double-click selection will continue to be specific to each type of content.

The user can pick what portion is shown in the middle level by dragging the indicator at the bottom of the top level. Alternatively, the scroll buttons at the edges of the middle level cause the viewed section to shift left and right. The scale of the video that is shown in the middle level can be adjusted by changing the size of the indicator at the bottom of the top level, by dragging on the edge of the indicator. This will zoom the middle level in and out. Similarly, the bottom level can be scrolled and zoomed using the indicator at the bottom of the middle level or the bottom's scroll arrows.

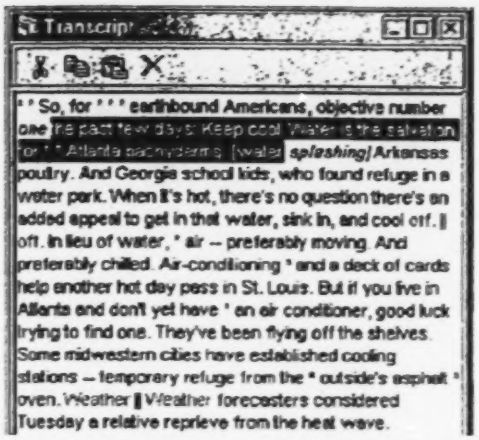
Transcript
" So, for " " earthbound Americans, objective number
one
splashing
Arkansas
poultry. And Georgia school kids, who found refuge in a
water park. When it's hot, there's no question there's an
added appeal to get in that water, sink in, and cool off. |
off. In lieu of water, " air -- preferably moving. And
preferably chilled. Air-conditioning " and a deck of cards
help another hot day pass in St. Louis. But if you live in
Atlanta and don't yet have " an air conditioner, good luck
trying to find one. They've been flying off the shelves.
Some midwestern cities have established cooling
stations -- temporary refuge from the " outside's asphalt "
oven. Weather | Weather forecasters considered
Tuesday a relative reprieve from the heat wave.

**Figure 4.** The Transcript view shows the text of the audio. The "*"s represent unrecognized portions of the audio, and the "|" represent breaks. The reverse video part at the top is selected. Words in italics are partially selected (here "splashing") and words in gray are partially cut out in the current production.

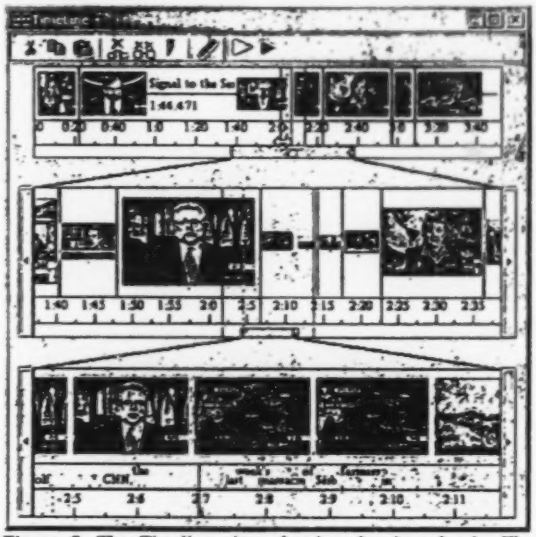

**Figure 5.** The Timeline view showing the three levels. The portion from about 2:2 to about 2:13 is selected, and appears on the screen in yellow in all three levels (and appears as light gray when printed in black-and-white).

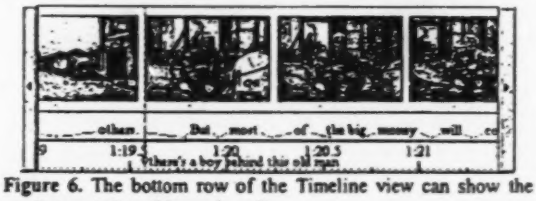

**Figure 6.** The bottom row of the Timeline view can show the user's annotations (shown in red).



BEST COPY AVAILABLE

The toolbar buttons at the top of the Timeline view window perform editing and playback operations (from left to right in Figure 5): cut, copy, paste, delete, crop (deletes everything but the selection), split (splices the clip at the selection edges), add annotation, play selection, and play the entire video.

Lee, Smeaton, *et. al.* [18] propose a taxonomy of video browsers based on three dimensions: the number of "layers" of abstraction and how they are related, the provision or omission of temporal information (varying from full timestamp information to nothing at all), and the visualization of spatial versus temporal aspects of the video (a slideshow is highly temporal, a timeline highly spatial). They recommend using many linked layers, providing temporal and absolute temporal information, and a spatial visualization. Our timeline follows their recommendations.

The Hierarchical Video Magnifier [24] and Swim [35] also provide multi-level views. These systems are designed to browse video, and navigation is achieved by drilling to higher levels of detail. The goal in Silver is to edit video and the basic interaction for navigating in the timeline is scrolling. Also, Silver is different in using multiple representations of the video within each level.

### 6.5 Preview View

As the user moves the cursor through the timeline, Silver displays a dotted line (visible to the left of 2.5 in all three levels of Figure 5). The frame at this point is shown in the preview view (Figure 7). If the user moves the cursor too fast, the preview view will catch up when the user stops or slows down sufficiently. The play arrows at the top of the timeline view cause the current video to be played in the preview view. If the black arrow is selected, the video is played in its entirety (but the user can always stop the playback). If the yellow play button is picked, only the selected portion of the video is played. The preview window is implemented using the Windows Media Player control.

### 6.6 Subject View

When creating a composition, different people have different ways of organizing their material. Some might like to group the material by topic. Silver's Subject View facilitates this type of organization. It provides a tabbed dialog box into which the material can be dragged-and-dropped from the project view. The user is free to label the tabs in any way that is useful, for example by the content of clip, the type of shot, the date, etc. The subject view (Figure 8) will allow the same clip to be entered multiple times, which will help users to more easily find material, since it might be classified in multiple ways.

**Figure 7.** The Preview view shows the frame at the cursor (the dotted lines in Figure 5), and is where the video is played.

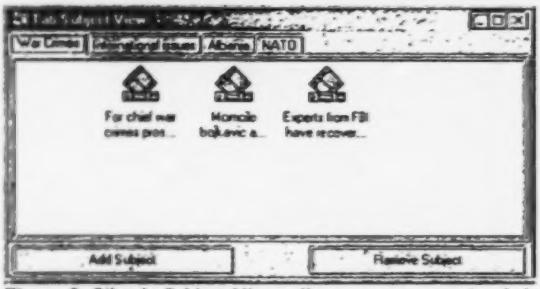


**Figure 8.** Silver's Subject Views allows users to organize their material by topic, type, date, etc.

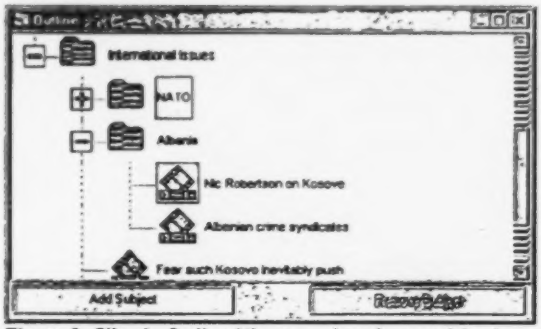


**Figure 9.** Silver's Outline View organizes the material using a Window's Tree control.

### 6.7 Outline View

When creating a composition, one good way to organize the material is in an outline. Whereas textual editing programs, such as Microsoft Word, have had outlining capabilities for years, none of the video editors have an outline view. Silver's outline view (shown in Figure 9) uses a conventional Windows tree control, which allows the hierarchy to be easily edited using familiar interaction techniques such as drag-and-drop. Note that for the subject view and the outline view, the subjects (or folders) can be added before there are any clips to put in them, to help guide the process and serve as reminders of what is left to do.



BEST COPY AVAILABLE

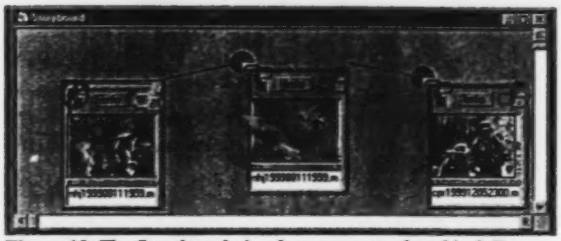

**Figure 10.** The Storyboard view has segments placed in 2-D.

### 6.8 Storyboard View

Many video and cinema projects start with a storyboard drawing of the composition, often drawn on paper. Typically, a picture in the storyboard represents each of the major scenes or sequences of the composition. Some video editors, notably MGI's VideoWave III, use a storyboard-like view as the main representation of the composition. Silver's storyboard view (see Figure 10) differs from VideoWave in that it can be used *before* the clips are found, as a representation of the desired video. Stills or even hand-drawn pictures can be used as placeholders in the storyboard for video to be shot or found later. The frames in the storyboard can be hand-placed in two dimensions by the user (and commands will help to visually organize them), which supports organizations that are meaningful to the user. For example, some productions are told using "parallel editing" [3] by cutting between two different stories occurring at the same time (for example, most *Star Wars* movies cut repeatedly between the story on a planet and the story in space). These might be represented in the storyboard by two parallel tracks.

Another important use for storyboards will be *interactive* video compositions (which Silver is planned to support in the future). Some multimedia productions allow the user to interact with the story using various methods to pick which video segment comes next. For example, a question might be asked or the user might click on various hot spots. Our storyboard view allows multiple arrows out of a clip, and we plan to support a "natural" scripting language [26] and demonstrational techniques [25] that will make it easy to specify how to choose which segment to play next based on the end user's input.

### 6.9 Other Views

In the future, we plan to add support for many other views, all inter-linked. For example, if the transcript window is used to hold an authored script, then it will be important to include "director's notes" and other annotations. These might be linked to views that help manage lists of locations, people, scenery, and to-do items. The ability to add notes, comments, annotations, and WWW links in all other views might also be useful. Other facilities from text documents might also be brought into the Silver editor, such as the ability to compare versions and keep track of changes (as a revision history).

### 6.10 Selection Across Multiple Views

When the user selects a portion of the video in one view in Silver, the equivalent portion is highlighted in all other views. This brings up a number of interesting user interface design challenges.

The first problem is what is the "equivalent" portion? The different views show different levels of granularity, so it may not be possible to represent the selection accurately in some views. For example, if a few frames are selected in the timeline view what should be shown in the project view since it only shows complete clips? Silver's design is to use dark yellow to highlight the selection, but to use light yellow to highlight an item that is only partially selected. In Figure 3, the first clip has only part of its contents selected, so it is shown in light yellow. If the user selects a clip in the project view, then all video that is derived from that clip is selected in all other views (which may result in discontinuous selections).

A similar problem arises between the timeline and transcript views. A particular word in the audio may span multiple frames. So selecting a word in the transcript will select all the corresponding frames. But selecting only one of those frames in the video may correspond to only part of a word, so the highlight in the transcript shows this by making that word *italic*. This is the case of the words "one" and "splashing" shown in the edges of the selected text in Figure 4. (We would prefer the selection to be yellow and the partially selected words in light yellow to be consistent with other views, but the Visual Basic text component does not support this.) If the selected video is moved, this will cut the word in two pieces. Silver represents this by repeating the word in both places, but showing it in a different font color. The video in Figure 4 was split somewhere during the word "Weather". Thus, this word is shown twice, separated by the clip boundary.

## 7. INTELLIGENT EDITING

One reason that video editing is so much more tedious than text editing is that in video, the user must select and operate on a frame-by-frame basis. Simply moving a section of video may require many minutes while the beginning and end points are laboriously located. Often a segment in the video does not exactly match with the corresponding audio. For example, the voice can start before the talking head is shown. This gives the video a continuous, seamless feel, but makes extracting pieces much harder because the video and audio portions must often be separately adjusted, with much fiddling to remove extraneous video or audio portions.

We did an informal study to check on the prevalence of these issues, and found it to indeed be significant, at least in recorded news. In looking at 238 transitions in 72 minutes of video clips recorded from CNN by Informedia, 61 (26%) were "L-cuts," where the audio and video come in or stop at different times. Of the 61 L-cuts, 44 (72%) were cases

where the audio of the speaker came in before the video but ended at the same time (see Figure 11-Shot A). Most of these were interviews where the voice of the person being interviewed would come in first and then the video of the person after a few seconds. Most of the other ways to overlap the video and audio were also represented. Sometimes, the audio and the video of the speaker came in at the same time, but the audio continued past the video (Figure 11-Shot B) or the video ended after the audio (Shot D). When an interviewee was being introduced while he appeared on screen, the video might come in before the audio, but both end at the same time (Shot C). To copy or delete any of these would require many steps in other editors.

The Silver editor aims to remove much of the tedium associated with editing such video by automatically adjusting the portions of the video and audio used for selection, cut, copy and paste, in the same way that text editors such as Microsoft Word adjust whether the spaces before and after words are selected. These capabilities are discussed in the following sections.

### 7.1 Intelligent Selection

When the user double-clicks in Microsoft Word and other text editors, the entire word is selected. Triple clicking will get the entire paragraph, sentence or line (depending on the editor). Silver provides a similar feature for video, and the unit selected on multiple clicks depends on which view is active. If the user double-clicks in the text view, the surrounding word or phrase will be selected that Informedia recognized (the minimal unit that can be matched with the audio). In the time-line view, however, the effect of double clicking depends on the specific timeline within the hierarchy. It can mean, for example, a shot (where shot boundaries are located automatically by Informedia) or an entire clip.

### 7.2 Intelligent Cut, Delete and Copy

When an operation is performed on the selected portion of the video, Silver uses heuristics to try to adjust the start and end points so they are appropriate for both the video and audio.

Using the information provided by Informedia, Silver will try to detect the L-cut situations, as shown in Figure 11. Silver will then try to adjust the selection accordingly. For example, when the user selects a sentence in the transcript and performs a copy operation, Silver will look at the corresponding video. If the shot boundary is not aligned with the selection from the audio, then Silver will try to determine if there is an L-cut as in Figure 11. If so, Silver will try to adjust the selected portion appropriately. However, editing operations using this selection will require special considerations, as discussed next.

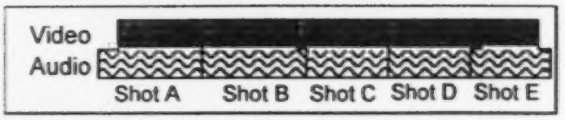


**Figure 11.** It is very common for the video and audio of a shot not to start and/or end at the same time. For example, Shot B represents a situation where the audio for the shot continues for a little while past when the video has already switched to the next scene. Similarly, for Shot D, the video continues a bit past when the audio has already switched to the next scene.

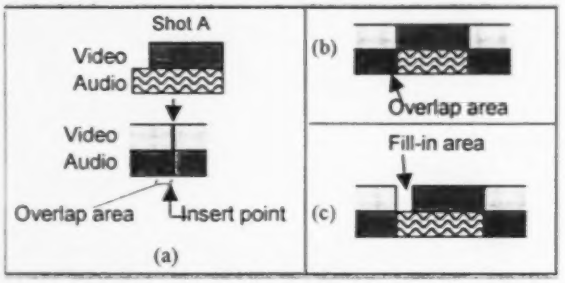


**Figure 12.** When an L-shot such as Shot A is inserted at a point in the video (a), Silver will check the area that might be overlapped. If the audio is silent in that area, Silver will overlap the audio automatically (b). In some cases (c), the user will have the option to overlay a separate piece of video if the audio cannot be overlapped.

### 7.3 Intelligent Paste and Reattach

When a segment with an L-cut is deleted or pasted in a new place, then Silver will need to determine how to deal with the uneven end(s). If the audio is shorter than the video (e.g., if Shots C or D from Figure 11 are pasted), then Silver can fill in with silence since this is not generally disruptive. However, it is not acceptable to fill in with black or blank video. For example, if Shot A is pasted, Silver will look at the overlapped area of the audio to see if it is silent or preceeded by an L-cut (Figure 12-a). If so, then the video can abut and the audio can be overlapped automatically (shown in Figure 12-b). If the audio in the overlap area is *not* silent, however, then Silver will suggest options to the user and ask the user which option is preferred. The choices include that the audio in the destination should be replaced, the audio should be mixed (which may be reasonable when one audio track is music or other background sounds), or else some video should be used to fill in the overlap area (as in Figure 12-c). The video to be filled in may come from the source of the copy (e.g., by expanding the video of Shot A) or else may be some other material or a "special effect" like a dissolve.

Although there are clearly some cases where the user will need to get involved in tweaking the edits, we feel that in the majority of cases, the system will be able to handle the edits automatically. It will await field trials, however, to measure how successful our heuristics will be.

## 8. INTELLIGENT CRITICS – FUTURE WORK

In school, children spend enormous amounts of time learning and practicing how to write. This includes learning the rules for organizing material, constructing sentences and paragraphs, and generally making a logical and understandable composition. However, few people will learn the corresponding rules for creating high-quality video and multimedia compositions. These are generally only taught in specialized, elective courses on film or video production.

Therefore, in order to help people create higher-quality productions, we plan to provide automatic critics that help evaluate and guide the quality of the production. Some of the techniques discussed in section 7 above will actually help improve the quality. We intend to go beyond this to provide many other heuristics that will watch the user's production and provide pop-up suggestions such as "avoid shaky footage" (as in [14]) and "avoid cutting in the middle of a camera pan."

Video editing is a highly subjective part of filmmaking, which can greatly affect the look of the finished product. Therefore, though some of the intelligent editing can be automated to prevent the user from making obvious errors, in some cases it is best to simply inform the user of the rule rather than make the artistic decision for them. For this reason, providing help as an Intelligent Critic is likely to be appropriate in this application.

A century of filmmaking has generated well-grounded theories and rules for film production and editing which can be used by our critic (e.g., [9] [21] [3]). For example, the effect of camera angle on comprehension and emotional impact [4], the effect of shot length and ordering on learning [13], and the effect of lighting on subjects' understanding of scenes [30], are just a small sample of film-making heuristics. As automatically generated metadata improves, it will be possible for Silver to give users more sophisticated assistance. For example, when Informedia vision systems are able to recognize similar scenes through an understanding of their semantic content, a future version of Silver could suggest that the first use of the scene be presented for a longer period than subsequent presentations. This is desirable to keep users' interest, keeping the user from becoming bored with the same visual material. Such capabilities in Silver will still not make Hollywood directors and editors of school children. However, it will provide a level of assistance that should enable naïve users to create much more pleasing productions from video archives.

## 9. CONCLUSIONS

We are implementing the Silver editor in Visual Basic, and although most functions are implemented, there is much more to be done. One important goal of the Silver project is to distribute our video editor so people can use it. Unfortunately, it is not yet robust enough to be put in front of users. It is also an important part of our plans to do informal and formal user tests, to evaluate and improve our designs.

As more and more video and film work is performed digitally, and as more and more homes and classrooms use computers, there will clearly be an increased demand for digital video editing. Digital libraries contain increasing amounts of multi-media material including video. Furthermore, people will increasingly want easy ways to put their own edited video on their personal web pages so it can be shared. Unfortunately, today's video editing tools make the editing of video significantly harder than the editing of textual material or still images. The Silver project is investigating some exciting ways to address this problem, and hopefully will point the way so the next generation of video editors will be significantly easier to use.

## ACKNOWLEDGEMENTS

The Silver Project is funded in part by the National Science Foundation under Grant No. IIS-9817527, as part of the Digital Library Initiative-2. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation. Thanks to Bernita Myers for help with this paper.

## REFERENCES

[1] Apple Computer, iMac. 2000. http://www.apple.com/imac/.

[2] Boreczky, J., et al. "An Interactive Comic Book Presentation for Exploring Video," in CHI 2000 Conference Proceedings. 2000. ACM Press. pp. 185-192.

[3] Cantine, J., Howard, S., and Lewis, B. Shot by Shot: A Practical Guide to Filmmaking, Second Edition. 1995, Pittsburgh Filmmakers.

[4] Carroll, J.M. and Bever, T.G. "Segmentation in Cinema Perception." Science, 1976. 191: pp. 1053-1055.

[5] Chiueh, T., et al. "Zodiac: A History-Based Interactive Video Authoring System," in Proceedings of ACM Multimedia '98. 1998. Bristol, England:

[6] Christel, M., et al. "Techniques for the Creation and Exploration of Digital Video Libraries," *Multimedia Tools and Applications*, B. Furht, Editor 1996, Kluwer Academic Publishers. Boston, MA.

[7] Christel, M., Winkler, D., and Taylor, C. "Multimedia Abstractions for a Digital Video Library," in Proceedings of the 2nd ACM International Conference on Digital Libraries. 1997. Philadelphia, PA: pp. 21-29.

[8] Chua, T. and Ruan, L., "A video retrieval and sequencing system." ACM Transactions on Information Systems, 1995. 13(4): pp. 373-407.

[9] Dancyger, K. The Technique of Film and Video Editing: Theory and Practice. Second ed. 1997, Boston, MA: Focal Press.

[10] Davenport, G., Smith, T.A., and Pincever, N. "Cinematic Primitives for Multimedia." IEEE Computer Graphics & Applications, 1991. 11(4): pp. 67-74.

[11] Fischhoff, B., et al. What Could You Do? Carnegie Mellon University, 1998. http://www.cmu.edu/telab/telfair_program/Fischhoff.html. Interactive Video DVD.

[12] Gauch, S., Li, W., and Gauch, J. "The VISION Digital Video Library." Information Processing & Management, 1997. 33(4): pp. 413-426.

[13] Gavriel, S. "Internalization of Filmic Schematic Operations in Interaction with Learners' Aptitudes." Journal of Educational Psychology, 1974. 66(4): pp. 499-511.

[14] Girgensohn, A., et al. "A semi-automatic approach to home video editing," in Proceedings of UIST'2000: The 13th annual ACM symposium on User interface software and technology. 2000. San Diego, CA: ACM. pp. 81-89.

[15] Hampapur, A., Jain, R., and Weymouth, T. "Digital Video Segmentation," in Proceedings of the Second ACM International Conference on Multimedia. 1994. San Francisco: pp. 357-364.

[16] Hauptmann, A. and Smith, M. "Text, Speech, and Vision for Video Segmentation: The Informedia Project," in AAAI Symposium on Computational Models for Integrating Language and Vision. 1995.

[17] Hauptmann, A.G. and Smith, M. "Video Segmentation in the Informedia Project," in IJCAI-95: Workshop on Intelligent Multimedia Information Retrieval. Montreal, 1995. Montreal, Quebec, Canada.

[18] Lee, H., et al. "Implementation and Analysis of Several Keyframe-Based Browsing Interfaces to Digital Video," in Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2000). 2000. Lisbon, Portugal: http://lorca.compapp.dcu.ie/Video/Papers/ECDL2000.pdf.

[19] Mackay, W.E. and Davenport, G. "Virtual Video Editing in Interactive Multimedia Applications." CACM, 1989. 32(7): pp. 832-843.

[20] Mackay, W.E. and Pagani, D. "Video Mosaic: Laying out time in a physical space," in Proceedings of the second ACM international conference on multimedia. 1994. San Francisco, CA: ACM. pp. 165-172.

[21] Mascelli, J. The Five C's of Cinematography: Motion Picture Filming Techiques. 1965, Los Angeles, CA: Silman-James Press.

[22] Matthews, J., Gloor, P., and Makedon, F. "VideoScheme: A Programmable Video Editing System for Automation and Media Recognition," in ACM Multimedia'93 Proceedings. 1993. pp. 419-426.

[23] Meng, J. and Chang, S. "CVEPS: A Compressed Video Editing and Parsing System," in Proceedings of the fourth ACM international conference on multimedia. 1996. Boston, MA: pp. 43-53.

[24] Mills, M., Cohen, J., and Wong, Y. "A Magnifier Tool for Video Data," in SIGCHI '92 Conference Proceedings of Human Factors in Computing Systems. 1992. Monterey, CA: ACM. pp. 93-98.

[25] Myers, B.A. "Demonstrational Interfaces: A Step Beyond Direct Manipulation." IEEE Computer, 1992. 25(8): pp. 61-73.

[26] Myers, B.A. Natural Programming: Project Overview and Proposal. Technical Report, Carnegie Mellon University School of Computer Science, CMU-CS-98-101 and CMU-HCII-98-100, January, 1998. Pittsburgh.

[27] Placeway, P., et al. "The 1996 Hub-4 Sphinx-3 System," in DARPA Spoken Systems Technology Workshop. 1997.

[28] Sato, T. et al. "Video OCR: Indexing Digital News Libraries by Recognition of Superimposed Caption." ACM Multimedia Systems: Special Issue on Video Libraries, 1999. 7(5): pp. 385-395.

[29] Stevens, S.M., Christel, M.G., and Wactlar, H.D. "Informedia: Improving Access to Digital Video." interactions: New Visions of Human-Computer Interaction, 1994. 1(4): pp. 67-71.

[30] Tannenbaum, P.H. and Fosdick, J.A. "The Effect of Lighting Angle on the Judgment of Photographed Subjects." AV Communication Review, 1960. pp. 253-262.

[31] Tonomura, Y., et al. "VideoMAP and VideoSpaceIcon: Tools for Anatomizing Video Content," in Proceedings INTERCHI'93: Human Factors in Computing Systems. 1993. ACM. pp. 131-136.

[32] Ueda, H. and Miyatake, T. "Automatic Scene Separation and Tree Structure GUI for Video Editing," in Proceedings of ACM Multimedia '96. 1996. Boston:

[33] Ueda, H., Tmiyatake, T., and Yoshizawa, S. "Impact: An Interactive Nautral-Motion-Picture Dedicated Multimedia Authoring System," in Proceedings of INTERCHI'93: Conference on human factors in computing systems. 1993. Amsterdam: ACM. pp. 137-141.

[34] Wactlar, H.D., et al. "Lessons learned from building a terabyte digital video library." IEEE Computer, 1999. 32(2): pp. 66 -73.

[35] Zhang, H., et al. "Video Parsing, Retrieval and Browsing: An Integrated and Content-Based Solution," in ACM Multimedia 95: Proceedings of the third ACM international conference on Multimedia. 1995. San Francisco, CA: pp. 15-24.